Claims

What is claimed is:

1. A system to facilitate building a statistical model for a data set, comprising:

a first training algorithm operative to efficiently build a first model for each subset of the data set;

an evaluation function operable to determine whether a subset of the data set for which the respective first model was built is an appropriate subset of the data set; and

a second training algorithm operable to build a second model based on the appropriate subset of the data set, the second training algorithm being more accurate than the first training algorithm.

2. The system of claim 1, further comprising a data scheduler which, based on a data policy, is operative to control the size of subsets for which the first training algorithm is applied.

3. The system of claim 2, wherein the data scheduler is operable to increase the size of the subset to provide a larger aggregate subset of the data set if the first model is unacceptable, the first training algorithm being operative to efficiently build the first model for each larger aggregate subset of the data until the evaluation function determines the resulting first model to be acceptable.

4. The system of claim 3, wherein the acceptability of each first model is determined based on a stopping criterion functionally related to an expected incremental benefit and a cost associated with increasing the size of the aggregate subset of the data set.

5. The system of claim 4, wherein the cost of the stopping criterion is functionally related to at least one of time associated with evaluating an aggregate data subset of increased size and size of the aggregated subset of the data.

6. The system of claim 4, wherein the stopping criterion is defined by

$$\left(\frac{l(D_{HO}\mid\theta(D_n))-l(D_{HO}\mid\theta(D_{n-1}))}{l(D_{HO}\mid\theta(D_n))-l(D_{HO}\mid\theta_{BASE}(D_n))}\right)\frac{1}{c_1(I_1-\bar{J}_n)\mid\Delta D_{n+1}\mid+c_2(I_1-\bar{J}_n)+c_1\bar{J}_n\mid D_{n+1}\mid+c_2\bar{J}_n+c_3}<\lambda$$

where

$l(D_{HO}|\theta(D_n))$ is a log likelihood for holdout data evaluated for the model built by the first training algorithm on a current subset of the training data set,

$l(D_{HO}|\theta(D_{n-1}))$ is a log likelihood for holdout data evaluated for the model built by the first training algorithm on a previous subset of the training data set,

$l(D_{HO}|\theta_{base}(D_n))$ is a log likelihood for holdout data evaluated for a base model,

c_1, c_2, and c_3 are constants determined based on application of the second training algorithm relative to a first subset of the data set,

I_1 is a number of iterations for the second training algorithm, when applied to the first subset,

$\bar{J}_n = \frac{1}{n}\sum_{i=1}^{n} J_i$, and

J_i is the number of iterations for the first training algorithm when applied to a data subset D_i,

$|D_{n+1}|$ is the size of data set D_{n+1},

$|\Delta D_{n+1}|$ is the increment in size $|D_{n+1}| - |D_n|$,

λ is a user determined stopping threshold .

7. The system of claim 4, wherein the stopping criterion is defined by

$$\left(\frac{l(D_{HO}\mid\theta(D_n))-l(D_{HO}\mid\theta(D_{n-1}))}{l(D_{HO}\mid\theta(D_n))+\delta-l(D_{HO}\mid\theta_{BASE}(D_n))}\right)\frac{1}{c_1(I_1-\bar{J}_n)\mid\Delta D_{n+1}\mid+c_2(I_1-\bar{J}_n)+c_1\bar{J}_n\mid D_{n+1}\mid+c_2\bar{J}_n+c_3}<\lambda$$

where

$l(D_{HO}|\theta(D_n))$ is a log likelihood for holdout data evaluated for the model built by the first training algorithm on a current subset of the training data set,

$l(D_{HO}|\theta(D_{n-1}))$ is a log likelihood for holdout data evaluated for the model built by the first training algorithm on a previous subset of the training data set,

$l(D_{HO}|\theta_{base}(D_n))$ is a log likelihood for holdout data evaluated for a base model,

δ is an offset associated with a difference in log likelihood for holdout data when evaluated for models built on a first subset of the training data set by the respective first and second training algorithms,

c_1, c_2, and c_3 are constants determined based on application of the second training algorithm relative to a first subset of the data set,

I_1 is a number of iterations for the second training algorithm, when applied to the first subset,

$$\bar{J}_n = \frac{1}{n}\sum_{i=1}^{n} J_i \text{, and}$$

J_1 is the number of iterations for the first training algorithm when applied to a data subset D_i,

$|D_{n+1}|$ is the size of data set D_{n+1},

$|\Delta D_{n+1}|$ is the increment in size $|D_{n+1}| - |D_n|$, and

λ is a user determined stopping threshold.

8. The system of claim 1, wherein the first training algorithm further comprises an iterative algorithm, which is operative to build the model for the subset of the data set according to an associated training policy.

9. The system of claim 8, wherein the first training algorithm further comprises an associated training policy that defines parameter initialization of the first training algorithm for each subset of the data set.

10. The system of claim 9, wherein the training policy associated with the first training algorithm further controls parameter initialization of the first training algorithm, such that at least some of the parameters computed for a previous subset of the data are employed to initialize the first training algorithm for a subsequent larger aggregate subset of the data.

11. The system of claim 9, wherein the first training algorithm is initialized by the same parameter values for each subset of the data subset.

12. The system of claim 9, wherein the training policy sets the iterative algorithm to perform a fixed number of at least one iteration.

13. The system of claim 12, wherein the training policy sets the iterative algorithm to perform a single iteration.

14. The system of claim 12, wherein the second training algorithm further comprises an iterative algorithm that operates according to an associated training policy, so as to produce a more accurate model for the appropriate subset of the data set than the first training algorithm.

15. The system of claim 14, wherein the iterative algorithm associated with at least one of the first and second training algorithms is an Expectation and Maximization algorithm.

16. The system of claim 8, wherein the training policy associated with the iterative algorithm of the first training algorithm controls the iterative algorithm to run until an associated convergence criterion is satisfied.

17. The system of claim 16, wherein second training algorithm further comprises an iterative algorithm, which is operative to build the model for the appropriate subset of the data set according to an associated training policy.

18. The system of claim 17, wherein the training policy associated with the iterative algorithm of the second training algorithm controls the respective iterative algorithm to run until an associated convergence criterion is satisfied, wherein the convergence criterion associated with the second training algorithm provides improved model quality relative to the convergence criterion associated with the first training algorithm.

19. A system programmed to facilitate building a statistical model, comprising:

a first parameter estimation algorithm operable to efficiently build models for subsets of a data set based on a training policy associated therewith; and

an evaluation function operable to determine whether a subset of data for which the model was built is an appropriate size for building the statistical model to characterize the data set;

a second parameter estimation algorithm operable to build a model on a subset of the data set having the appropriate size, the second parameter estimation algorithm having an associated training policy, which enables the second parameter estimation algorithm to build a more accurate model than the first parameter estimation algorithm.

20. The system of claim 19, further comprising a data scheduler operable to increase the size of the subset of the data set to provide a larger aggregate subset of the data set if the model is unacceptable, the first parameter estimation algorithm being operative to efficiently build a model for each larger aggregate subset until a resulting model built therefrom is determined to be acceptable.

21. The system of claim 19, wherein the first parameter estimation algorithm further comprises an iterative algorithm operative to build the model for each subset of the data set according to the associated training policy.

22. The system of claim 21, wherein the training policy for the first parameter estimation algorithm is operative to control parameter initialization for the first parameter estimation algorithm, such that at least some of the parameters computed for a previous subset of the data are employed to initialize the first parameter estimation algorithm for a subsequent larger aggregate subset of the data set.

23. The system of claim 21, wherein the first parameter estimation algorithm is initialized by the same parameter values for each subset of the data subset.

24. The system of claim 21, wherein the training policy associated with first parameter estimation algorithm controls the iterative algorithm of the first parameter estimation algorithm to perform a fixed number of at least one iteration, the second training algorithm further comprising an iterative algorithm, which is operative to perform a greater number of iterations than the iterative algorithm of the first training algorithm based on a training policy associated with the second parameter estimation algorithm.

25. The system of claim 21, wherein the training policy associated with the iterative algorithm of the first parameter estimation algorithm controls the iterative algorithm to run until an associated convergence threshold is satisfied, wherein the second training algorithm further comprises an iterative algorithm, the training policy associated with the iterative algorithm of the second parameter estimation algorithm being operative to control the respective iterative algorithm to run until an associated convergence threshold is satisfied, the convergence threshold associated with the second parameter estimation algorithm is less than the convergence threshold associated with the first parameter estimation algorithm.

26. The system of claim 19, wherein the evaluation function determines whether the subset of data for which the model was built is an appropriate size based on a stopping criterion, which is functionally related to an expected incremental benefit and an expected incremental cost associated with increasing size of the subset of data.

27. The system of claim 26, wherein the cost of the stopping criterion is functionally related to at least one of time associated with evaluating the model for a larger subset of data and size of the larger subset of the data.

28. The system of claim 26, wherein the stopping criterion is defined by

$$\left(\frac{l(D_{HO}\mid\theta(D_n))-l(D_{HO}\mid\theta(D_{n-1}))}{l(D_{HO}\mid\theta(D_n))-l(D_{HO}\mid\theta_{BASE}(D_n))}\right)\frac{1}{c_1(I_1-\bar{J}_n)\mid\Delta D_{n+1}\mid+c_2(I_1-\bar{J}_n)+c_1\bar{J}_n\mid D_{n+1}\mid+c_2\bar{J}_n+c_3}<\lambda$$

where

$l(D_{HO}|\theta(D_n))$ is a log likelihood for holdout data evaluated for the model built by the first training algorithm on a current subset of the training data set,

$l(D_{HO}|\theta(D_{n-1}))$ is a log likelihood for holdout data evaluated for the model built by the first training algorithm on a previous subset of the training data set,

$l(D_{HO}|\theta_{base}(D_n))$ is a log likelihood for holdout data evaluated for a base model,

c_1, c_2, and c_3 are constants determined based on application of the second parameter estimation algorithm relative to a first subset of the data set,

I_1 is a number of iterations for the second parameter estimation algorithm, when applied to the first subset,

$\bar{J}_n = \frac{1}{n}\sum_{i=1}^{n} J_i$, and

J_i is the number of iterations for the first parameter estimation algorithm when applied to a data subset D_i,

$|D_{n+1}|$ is the size of data set D_{n+1},

$|\Delta D_{n+1}|$ is the increment in size $|D_{n+1}| - |D_n|$, and

λ is a user determined stopping threshold.

29. The system of claim 26, wherein the stopping criterion is defined by

$$\left(\frac{l(D_{HO}\,|\,\theta(D_n)) - l(D_{HO}\,|\,\theta(D_{n-1}))}{l(D_{HO}\,|\,\theta(D_n)) + \delta - l(D_{HO}\,|\,\theta_{BASE}(D_n))}\right)\frac{1}{c_1(I_1-\bar{J}_n)\,|\,\Delta D_{n+1}\,| + c_2(I_1-\bar{J}_n) + c_1\bar{J}_n\,|\,D_{n+1}\,| + c_2\bar{J}_n + c_3} < \lambda$$

where

$l(D_{HO}|\theta(D_n))$ is a log likelihood for holdout data evaluated for the model built by the first training algorithm on a current subset of the training data set,

$l(D_{HO}|\theta(D_{n-1}))$ is a log likelihood for holdout data evaluated for the model built by the first training algorithm on a previous subset of the training data set,

$l(D_{HO}|\theta_{base}(D_n))$ is a log likelihood for holdout data evaluated for a base model,

δ is an offset associated with a difference in log likelihood for holdout data when evaluated for models built on a first subset of the training data set by the respective first and second training algorithms,

c_1, c_2, and c_3 are constants determined based on application of the second parameter estimation algorithm relative to a first data subset of the data set,

I_1 is a number of iterations for the second parameter estimation algorithm, when applied to a first data subset,

$\bar{J}_n = \frac{1}{n}\sum_{i=1}^{n} J_i$, and

J_i is the number of iterations for the first parameter estimation algorithm when applied to a data subset D_i,

$|D_{n+1}|$ is the size of data set D_{n+1},

$|\Delta D_{n+1}|$ is the increment in size $|D_{n+1}| - |D_n|$, and

λ is a user determined stopping threshold.

30. A learning curve method to facilitate building a statistical model, comprising:

choosing a subset of a data set;

employing a first training algorithm to build a first model to characterize the subset;

evaluating the first model;

if the first model is unacceptable, repeatedly increasing the size of the subset of data to provide an aggregate data set, building another first model to characterize the aggregate subset, and reevaluating the model; and

if the model is acceptable, employing a second training algorithm to build a second model based on the aggregate data set, the second training algorithm being different from the first training algorithm.

31. The method of claim 30, further comprising determining the acceptability of each first model based on a stopping criterion functionally related to an expected incremental benefit and an expected incremental cost associated with increasing the size of the aggregate subset of the data set.

32. The system of claim 31, wherein the cost of the stopping criterion is functionally related to at least one of time associated with evaluating an aggregate data subset of increased size and size of the aggregate subset of the data.

33. The system of claim 31, wherein the stopping criterion is defined by

$$\left(\frac{l(D_{HO}\mid\theta(D_n))-l(D_{HO}\mid\theta(D_{n-1}))}{l(D_{HO}\mid\theta(D_n))-l(D_{HO}\mid\theta_{BASE}(D_n))}\right)\frac{1}{c_1(I_1-\bar{J}_n)\mid\Delta D_{n+1}\mid+c_2(I_1-\bar{J}_n)+c_1\bar{J}_n\mid D_{n+1}\mid+c_2\bar{J}_n+c_3}<\lambda$$

where

$l(D_{HO}|\theta(D_n))$ is a log likelihood for holdout data evaluated for the model built by the first training algorithm on a current subset of the training data set,

$l(D_{HO}|\theta(D_{n-1}))$ is a log likelihood for holdout data evaluated for the model built by the first training algorithm on a previous subset of the training data set,

$l(D_{HO}|\theta_{base}(D_n))$ is a log likelihood for holdout data evaluated for a base model,

c_1, c_2, and c_3 are constants determined based on application of the second parameter estimation algorithm relative to a first subset of the data set,

I_1 is a number of iterations for the second parameter estimation algorithm, when applied to the first subset,

$$\bar{J}_n=\frac{1}{n}\sum_{i=1}^{n}J_i\text{, and}$$

J_i is a number of iterations for the first parameter estimation algorithm when applied to a data subset D_i,

$|D_{n+1}|$ is a size of data set D_{n+1},

$|\Delta D_{n+1}|$ is an increment in size $|D_{n+1}| - |D_n|$, and

λ is a user determined stopping threshold.

34. The system of claim 31, wherein the stopping criterion is defined by

$$\left(\frac{l(D_{HO}\mid\theta(D_n))-l(D_{HO}\mid\theta(D_{n-1}))}{l(D_{HO}\mid\theta(D_n))+\delta-l(D_{HO}\mid\theta_{BASE}(D_n))}\right)\frac{1}{c_1(I_1-\bar{J}_n)\mid\Delta D_{n+1}\mid+c_2(I_1-\bar{J}_n)+c_1\bar{J}_n\mid D_{n+1}\mid+c_2\bar{J}_n+c_3}<\lambda$$

where

$l(D_{HO}|\theta(D_n))$ is a log likelihood for holdout data evaluated for the model built by the first training algorithm on a current subset of the training data set,

$l(D_{HO}|\theta(D_{n-1}))$ is a log likelihood for holdout data evaluated for the model built by the first training algorithm on a previous subset of the training data set,

$l(D_{HO}|\theta_{base}(D_n))$ is a log likelihood for holdout data evaluated for a base model,

δ is an offset associated with the difference in log likelihood for holdout data when evaluated for models built on a first subset of the training data set by the respective first and second training algorithms,

c_1, c_2, and c_3 are constants determined based on application of the second parameter estimation algorithm relative to a first data subset of the data set,

I_1 is a number of iterations for the second parameter estimation algorithm, when applied to a first data subset,

$$\bar{J}_n = \frac{1}{n}\sum_{i=1}^{n} J_i \text{, and}$$

J_i is a number of iterations for the first parameter estimation algorithm when applied to a data subset D_i,

$|D_{n+1}|$ is a size of data set D_{n+1},

$|\Delta D_{n+1}|$ is an increment in size $|D_{n+1}| - |D_n|$, and

λ is a user determined stopping threshold.

35. The method of claim 30, wherein the first training algorithm is more computationally efficient than the second training algorithm.

36. The method of claim 30, wherein each instance of model building repeated until obtaining an acceptable model by the first training algorithm employs more efficient and less accurate model building than model building employed by the second training algorithm that occurs after obtaining the acceptable model.

37. The method of claim 36, wherein each instance of model building repeated until obtaining an acceptable model employs the first training algorithm as an iterative algorithm that is run to a first convergence criterion, the second training algorithm employing an iterative algorithm that is run to a second convergence criterion, which demands more

iterations than the first convergence criterion in order to obtain convergence, so that the second model is more accurate than the first model built by the first training algorithm.

38. The method of claim 36, wherein each instance of model building repeated until obtaining an acceptable model employs an iterative algorithm having a fixed number of at least one iteration, the second training algorithm employing an iterative algorithm having a greater number of iterations than the fixed number.

39. The method of claim 30, further comprising controlling parameter initialization employed in each instance of building a model for the aggregate data set prior to obtaining an acceptable model.

40. The method of claim 39, further comprising initializing the first training algorithm by the same parameter values for each subset.

41. The method of claim 39, wherein the controlling further comprises reusing at least some of the parameters computed from a previous instance of model building to initialize a subsequent instance of model building for a subsequent larger aggregate data set prior to obtaining an acceptable model.

42. A computer-readable medium having computer-executable instructions for:

choosing a subset of a data set;

building a model to characterize the subset based on an associated training policy;

evaluating the model;

if the model is unacceptable, repeatedly increasing the size of the subset of data to provide an aggregate data set, building a model to characterize the aggregate subset based on an associated training policy, and reevaluating the model; and

if the model is acceptable, employing the aggregate data set to build a corresponding model based on an associated training policy, the training policy associated with the model building repeated until obtaining an acceptable model being more

computationally efficient than the training policy associated with model building subsequent thereto.

43. The method of claim 42, further comprising determining the acceptability of the model based on an expected incremental benefit relative to an expected incremental cost associated with increasing the size of the aggregate data set.

44. A method to facilitate constructing a statistical model, comprising:

separating data into holdout data and training data;

determining a data subset from the training data by estimating model parameters according to a first training policy and evaluating the estimated model parameters relative to the holdout data set and repeating the estimation and evaluation of model parameters with a larger subset of the training data until an acceptable quality of the estimated model is established; and,

subsequent to establishing the acceptable quality of the estimated model, using the determined data subset to improve the estimated model parameters by employing a second training policy that is more accurate than the first training policy.

45. The method of claim 44, wherein each estimation of model parameters repeated until the acceptable quality of the estimated model is established further comprises employing an iterative algorithm that is run until a first convergence criterion is satisfied, the estimation of model parameters using the determined data subset further comprising an iterative algorithm that is run until a second convergence criterion is satisfied, which is operative to provide a better quality of model than the first convergence criterion.

46. The system of claim 45, wherein the first convergence criterion causes the associated iterative algorithm to run until a first convergence threshold is satisfied, wherein the second convergence criterion causes the associated iterative algorithm to run until a second convergence threshold is satisfied, the second convergence threshold being less than the first convergence threshold.

47. The method of claim 45, wherein at least one of the iterative algorithm run to the first convergence criterion and the iterative algorithm run to the second convergence criterion is an Expectation and Maximization algorithm.

48. The method of claim 44, wherein each estimation of model parameters repeated until the acceptable quality of the estimated model is established employs an iterative algorithm having a fixed number of at least one iteration, the estimation of model parameters using the determined data subset further employing an iterative algorithm having a greater number of iterations than the fixed number.

49. The method of claim 44, further comprising controlling parameter initialization employed in each estimation of model parameters repeated until determining an acceptable size for the determined data subset.

50. The method of claim 44, wherein the controlling further comprises reusing at least some of the parameters computed from a previous estimation of model parameters to initialize a subsequent estimation of model parameters for a next larger subset of the training set.

51. The method of claim 44, wherein each estimation of model parameters repeated until the acceptable quality of the estimated model is established further comprises initializing the first training algorithm by the same parameter values.

52. The method of claim 44, further comprising determining the acceptability of the estimated model based on an expected incremental benefit relative to a cost associated with increasing the size of the subset of the data set.

53. A computer-readable medium having computer-executable instructions for:

separating data into holdout data and training data;

determining a data subset from the training data by estimating model parameters according to a first training policy and evaluating the estimated model parameters

relative to the holdout data set and repeating the estimation and evaluation of model parameters with a next successively larger subset of the training data set until an acceptable quality of the estimated model is established; and

subsequent to establishing the acceptable quality of the estimated model, using the determined data subset to improve the estimated model parameters by employing a second training policy that is more accurate than the first training policy.

54. A method to facilitate constructing a statistical model, comprising:

separating data into a holdout data set and a training data set;

iteratively estimating model parameters for a subset of the training data set over a fixed number of iterations and evaluating the estimated model parameters relative to the holdout data set;

repeating the estimation and evaluation of model parameters obtained with successively larger subsets of the training data set until an acceptable model quality is established; and

after the acceptable model quality is established, iteratively estimating model parameters for the data subset, which provided the acceptable model quality, until a better quality of model is provided relative to a preceding estimation performed over the fixed number of iterations.

55. The method of claim 54, wherein at least one of the iterative estimations employs an Expectation and Maximization algorithm.

56. The method of claim 54, wherein the estimation that occurs after the acceptable model quality is established, further comprises employing an iterative algorithm having a greater number of iterations than the fixed number.

57. The method of claim 54, wherein the estimation of model parameters after the acceptable model quality has been established further comprises employing an iterative algorithm that is run until a convergence criterion is satisfied, which is operative to provide a

better quality of model with the data subset than a preceding estimation employing the fixed number of iterations.

58. The method of claim 54, further comprising controlling parameter initialization for each estimation of model parameters that occurs before the acceptable model quality has been established.

59. The method of claim 58, wherein each iterative estimation until the acceptable model quality is established further comprises initializing the first training algorithm by the same parameter values.

60. The method of claim 58, wherein the controlling further comprises reusing at least some of the parameters obtained in a previous estimation of model parameters to initialize a subsequent estimation of model parameters for a next larger subset of the training data set.

61. The method of claim 54, further comprising determining the acceptability of the model based on an expected incremental benefit relative to an expected incremental cost associated with an increase in size of each larger training subset of the data set.

62. A method to facilitate constructing a statistical model, comprising:

separating data into a holdout data set and a training data set;

iteratively estimating model parameters for a subset of the training data set until a first convergence threshold is satisfied and evaluating the estimated model parameters relative to the holdout data set;

repeating the estimation and evaluation of model parameters obtained with successively larger subsets of the training data set until determining a size of data subset that provides acceptable model parameters; and

after determining the size of data subset that provides acceptable model parameters, iteratively estimating model parameters for a data subset of the acceptable size

until a second convergence threshold is satisfied, the second convergence threshold being less than the first convergence threshold.

63. A system to facilitate building a statistical model for a data set, comprising:

first means for building a first model to characterize a subset of the data set;

evaluation means for evaluating the acceptability of the model, the first means building another first model for a larger subset of the data if the evaluation means determines that a prior first model is unacceptable; and

second means, which is different from the first means, for building a second model to characterize an aggregate subset of data that enabled the first means to produce an acceptable model.

64. A system to facilitate building a statistical model for a data set, comprising:

first means for estimating model parameters from a subset of the data set ;

means for evaluating the estimated model parameters relative to a holdout set of the data set;

means for determining a data subset from the training data by causing the first means and the means for evaluating to respectively repeat estimation and evaluation of model parameters with a next successively larger subset of the training data set until an acceptable quality of the model parameters is established; and

second means for estimating model parameters based on the determined data subset to provide a more accurate estimation of model parameters than the first means.